DOCUMENT RESUME

ED 254 868 CS 504 814

AUTHOR Crocker-Lakness, Jim
TITLE The Accuracy of Peer Judgment of Debates in the Classroom.
PUB DATE 84
NOTE 7p.
PUB TYPE Reports - Research/Technical (143) -- Journal Articles (080)
JOURNAL CIT Kentucky Journal of Communication Arts; v10 n1 p6-9 Fall 1984

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Communication Research; Comparative Analysis; *Debate; *Egocentrism; *Evaluation Criteria; Higher Education; Influences; *Judges; *Peer Evaluation; Public Speaking; Speech Communication

ABSTRACT

Data were collected over a three-year period from 10 sections of an advanced public speaking course at a midwestern university in a study to determine (1) the extent to which debaters who are not ego-involved can accurately judge classroom debate speeches, and (2) the specific variables that affect this accuracy. All 10 sections were taught by the same instructor/judge. As part of the course, each student was required to speak in two debates, once as an affirmative speaker and once as a negative speaker. The course instructor/judge and two student debaters served as judges to evaluate each debate. The five-point rating scale (from poor to superior) for each of the six judgment criteria (analysis, reasoning, evidence, organization, refutation, and delivery) was retained as was the requirement that judges decide whether the affirmative or negative won the debate. Results showed that less ego-involved debaters showed significant moderate correlations with the judge's ratings. But the correlations were not high enough in comparison with other studies to conclude that debaters' judgments markedly improved when they were not ego-involved in the outcome of the debate or when they were judging classroom rather than tournament debates. Correlations of overall and specific criteria ratings were higher for negative debate speeches than for affirmative. The study concluded that student debaters did not have sufficient higher level cognitive awareness to enable them to evaluate the quality of a debate. (HOD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED254868

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

The Accuracy of Peer Judgment of Debates in the Classroom

Jim Crocker-Lakness

CS 504 814

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Jim Crocker-Lakness

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The Accuracy of Peer Judgment of Debates in the Classroom

*Jim Crocker-Lakness

The question of how accurately debaters judge debates has been asked repeatedly since Baccus'[1] first report on the issue. For some the answer has obvious implications for tournament judging. If debaters can accurately judge debates, their judgments could be employed as a check on judgments by the critic judges. A more far reaching implication is for debates pedagogy. Accurate judgments by debaters would suggest the existence of some objective criteria by which quality debates and debaters are recognized by critic judges and debaters alike. Identifying these criteria, if they exist, would make debate pedagogy more specific and concrete. If debaters themselves are able accurately to judge debates, then self-help practice sessions could supplement direct supervision and evaluation by the instructor coach. To the extent training in debate teaches objective critical listening, we could reasonably expect debaters to make good judges of debate. If debaters are unable to judge debates accurately, the value of debate training to teach critical listening and higher order cognitive skills may be called into question. Bloom's[2] taxonomy of learning places application and evaluation at the top of his hierarchy of cognitive learning goals. It is entirely possible for a student to be an excellent debater and yet not understand how or why. A debater's ability *to judge* quality debates as well as present them is one measure of the debater's skill at Bloom's higher levels of learning. If debaters cannot accurately judge debates, it may be that debate training does not adequately lead a student to these goals.

If this is the case, identifying how debaters and judges differ in their judgments can be useful information for coaches and instructors alike. Coaches will want to help debaters have a more accurate assessment of debates, and instructors need to identify specific areas in which student debater's judgements can be improved. Identifying these areas may shed light on the kind of critical judgment student debaters learn.

Previous Research

This study was designed to continue the research begun by others on this question. Several studies have contributed to an answer, but the evidence to date is incomplete and so inconclusive.

After the first three rounds of an intercollegiate debate tournament, King and Cleavenger[3] asked participant debaters to rank order their opponents. When these rank orders were compared to judges' rank orders, significant positive correlations were found. King and Cleavenger did not collect interval data or data from specific evaluation criteria. While they concluded that the outcome of the tournament would have been different if the participants rather than the judges had made decisions, their results do not directly shed much insight on debate pedagogy.

In his correlation study Barker concluded that "debaters and judges do not rate debate performance alike"[4] regardless of the debater's level of experience or debate side (affirmative or negative). His results led him to speculate that debaters and judges have different concepts of the criteria of evaluation and to cast doubt on the objective critical thinking abilities of debaters. Two procedures, however, limit the generalizability of his results. He asked debaters to evaluate their *own* debates as a measure of debater judging accuracy. Thus debaters are probably ego-involved in their judgements, and their ego involvement could well be a confounding variable. In addition, Barker's subjects came for both tournament and classroom samples, confusing the generalizability of his results. The only significant correlation Barker found was for his inexperienced (i.e., classroom rather than tournament) sample where the criteria of the instructor judge could assumed to be more obvious to the debater than in the experienced sample. This correlation, however, was based on such a small sample that Barker discounted it. While suggesting, then, that tournament debaters do not accurately judge their own debates, he left open the question whether debaters in nontournament settings can accurately judge the debates of others.

Hill[5] did draw his sample entirely from the classroom and asked debaters to judge the debates of others as well as their own. Each debater was asked to evaluate his or her own debate, that of the colleague, and those of opponents. The only significant correlations Hill found were between a classroom debater's evaluation and a judge's evaluation and between a debater's evaluation by his or her debate colleague and a judge's evaluation. His explanation of the difference between his results and Barker's was that specific criteria for evaluation were more obvious to debaters in the classroom than in a tournament. Since, however, Hill asked his debaters only to judge debates in which they were involved, he did not eliminate ego involvement from his sample.

The results of previous research yield some conclusions and more questions. Apparently, although no agreement was found between debaters' self evaluations and judges' evaluations in tournaments, debaters in the classroom are somewhat able to judge their own debates though not those of opponents. The issue of ego involvement is unresolved and the question posed earlier for pedagogy left unanswered. Hill's recommendation for "further research with subjects who are not involved in intercollegiate debate"[6] deserves attention, and Barker's suggestion that "practice in judging the debates of others in his class would be a practical way to

*Jim Crocker-Lakness is an Assistant Professor in the Department of Speech, University of Cincinnati.

let the debater discover what factors to look for in evaluating debates"[7] is worthy of test.

Furthermore, a closer examination of previous work suggests that the debate side being evaluated (affirmative or negative) may be a variable in the degree of accuracy of a debater's judgement. King and Cleavenger found a higher correlation (r=.843) for participant and judge rank orders of negative speeches than for similar rankings of affirmative speeches (r=.496). Both Barker and Hill performed separate analyses of judgements on affirmative and negative sides. While Barker found a positive correlation with judges' decisions for inexperienced affirmative speakers, the decisions of corresponding negative debaters showed no significant correlation. From his results Hill points out that self evaluations by first affirmative speakers correlate higher with judges' evaluations than do speeches in other positions and calls for further research to explain this phenomenon. He does not notice that the debate side of the speaker being evaluated may affect the correlations found. Of his correlations for 32 sub-sets of all combinations of judge and four speakers evaluating each speaker in turn, 11 correlations were significant when the speaker being evaluated was the negative side while only 4 correlations are significant when the affirmative speaker was the object of judgement.

The present study was designed to explore these areas suggested but untested by previous research. Two research problems are addressed: (1) To what extent can debaters who are not ego involved accurately judge classroom debate speeches? and (2) What specific variables affect this accuracy ? In this study as in previous work, accuracy by student debaters is defined as agreement with critic judges.

For statistical purposes four null hypotheses were posited:

1a) Debaters in the classroom will not accurately judge which side won the debate.

1b) Debaters will be no more accurate in judging that the affirmative side won than in judging that the negative side won.

2a) Considering affirmative and negative speeches separately, there will be no significant correlations between overall ratings of debate speeches by debaters and an instructor judge.

2b) Considering affirmative and negative speeches separately, there will be no significant correlations between debaters' and an instructor/judge's ratings of debate speeches on specific rating criteria.

3) There will be no differences between debater/judge correlations on specific criteria for affirmative speeches and debater/judge correlations on specific criteria for negative debate speeches.

4) For affirmative and negative speeches considered separately, there will be no differences between mean ratings on specific criteris by debaters and an instructor judge.

Procedures

Decisions on who won a debate and ratings of debates by an instructor/judge and by classroom debaters were compared. Data was collected over a three year period from ten sections of an Advanced Public Speaking course at a midwestern state university. All ten sections are taught by the same instructor/judge. As part of this course each student was required to speak in two debates, once as an affirmative speaker and once as a negative speaker. Debates adhered to a modified format for one debater on each side rather than two. The affirmative speaker spoke for five minutes followed by a negative presentation for seven minutes which was in turn followed by the affirmative rebuttal for two minutes.

The course instructor/judge and two student debaters served as judges to evaluate each debate. Student judges rotated on a fixed schedule so that student judges during any class meeting were those students scheduled to debate during the next class meeting. The last row of seats in the classroom was reserved for the judges who sat separated from each other by enough distance to prevent them from seeing how each other rated individual debates and debaters. After each debate judges announced their decisions, and the rating ballots were placed in a folder. The instructor did not examine ballots by student judges to avoid a possible bias in the instructor/judge's future ratings toward or away from agreement with student ratings. Students saw only the instructor's ratings of their own speeches, never ratings of each other's speeches.

Judging decisions were recorded on an AFA Form C ballot modified to accommodate two speakers per debate rather than four. The five point rating scale (from poor to superior) for each of the six judgment criteria (analysis, reasoning, evidence, organization, refutation and delivery) was retained as was the requirement that judges decide whether affirmative or negative won the debate. The meanings of these six rating dimensions and the use of the rating scale instrument were explained to students as part of the course. While the reliability of this rating instrument was been called into question,[8] some evidence shows that debaters prefer it as a learning tool over other rating forms.[9].

Using this method, data was collected from 109 debates. Because of incomplete data or inaccurate use ballots from 9 debates were eliminated leaving data from a total of N = 100 debates for analysis. This data consisted of instructor/judge and student ratings for each affirmative and negative speaker on each of the six criteria and a composite score, defined as the sum of the ratings on the six criteria. The vote of each judge stating who won each debate were also collected.

Rating data was analyzed by the product moment correlation. The average of the two students' ratings for each debate were compared to the instructor/judge's ratings. Correlations were computed for composite ratings and for ratings on each of the six specific criteria. Decisions on which side won the debate were analyzed by a 2 x 3 Chi-square with the type of speech judged (affirmative or negative) as column entries and the kind of agreement (unanimous, both debator judges disagreeing with the instructor, and ambiguous 2-1 splits of three judges) as row entries. A probability level of .05 was selected as necessary for rejection of the null hypothesis. Debater and instructor/judge mean ratings on specific evaluation criteria were compared by a t-test.

Results

Hypothesis #1

Table 1 shows the distribution of decisions on who won the debate. Unanimous agreement between debaters and the

judge is more frequent (both when affirmative won and when negative won the debate) than either 2-1 splits by debaters against the instructor or ambiguous splits in which one debater agreed with the judge and the other debater did not. Unanimous agreement that the affirmative side won the debate is more frequent than unanimous agreement that the negative side won the debate. This would suggest that the side of the speech acts as a variable affecting the accuracy of debaters' judgments except for the fact that the Chi-square analysis failed to show significance and the required level. The null hypothesis is retained.

Table 1

Distribution of the Frequency of Decisions on Which Side Won the Debate by Two Debaters and an Instructor/Judge

	Unanimous	Decision Split Both Debaters Disagreeing with Instructor/Judge	Ambiguous
For the Affirmative	30	11	16
For the Negative	19	6	18

$x^2 = 2.1402$ not significant

Hypothesis #2

Correlations of instructor/judge and debaters' ratings are shown in Table 2. All are positive and are significant at the .001 level. All correlations for the negative speeches and those for the composite, evidence, organization and delivery ratings for the affirmative speeches are in the moderate range. The null hypothesis is rejected with reservations. While the correlations for composite ratings are higher than those found by either Barker or Hill, they are not sufficiently higher to warrant a conclusion that debaters not ego involved will judge debaters in the classroom more accurately than ego involved tournament debaters. The data cannot be directly compared because of methodological differences in collecting it. Barker's suggestion that debaters could improve their accuracy by judging the debates of others in the class is not exactly borne out.

Of the correlations for specific judgement criteria the highest for both affirmative and negative exists for the evidence criterion while the lowest are for analysis and reasoning. This result may be due to the fact that while analysis and reasoning are more abstract processes, the use of evidence may be more concrete, or at least thought to be. A citation of evidence may be more obvious to an inexperienced debater than the more abstract reasoning and analysis. Evidence can be judged quantitatively while assessment of analysis and reasoning require more qualitative decisions. This result suggests that debaters may not conceptualize the processes of debate at higher levels of cognitive learning.

Table 2

Correlations Between Ratings by the Instructor/Judge and the Average Rating by Two Debaters

	For Affirmative Speeches	For Negative Speeches
Composite	r = .53	r = .61
Analysis	.38	.40
Reasoning	.35	.40
Evidence	.51	.60
Organization	.40	.40
Refutation	.37	.55
Delivery	.49	.47

All correlations significant $p < .001$

Hypothesis #3

An examination of Table 2 shows that for every case except for delivery, the correlations for judgements of negative speeches are higher than those for affirmative speeches. The null hypothesis is rejected with reservations. Since this conclusion is supported by some results of earlier studies, it deserves some attention. Exactly why debaters tend to judge the negative side more accurately than the affirmative side is not exactly clear. A possible explanation resulting from the methodology is that since the negative speech follows the affirmative the debater's rating skill improves. Perhaps student debaters are more accustomed to hearing and making "negative" criticisms so that they are more familiar with it and thus can judge it more accurately. An explanation more related to the nature of the affirmative or negative speech is that negative debate speeches may be more consistent in quality. The task of the negative side in a debate may be easier than the affirmative task. Certainly the affirmative speaker must define terms, introduce and defend issues to make a case. The affirmative carries the burden. The negative speaker reacts rather than acts. It may be that because more is required of the affirmative speaker, finer evaluative distinctions can be made, at least by the instructor/judge. Whatever the explanation, apparently debaters in this sample were able to judge negative speeches somewhat more accurately than affirmative speeches.

Hypothesis #4

Rating means by the instructor/judge and by the debaters on specific criteria for affirmative and negative debate speeches are seen in Table 3. Two signigicant differences were found for evidence and refutation by negative speakers. The null hypothesis is rejected. It is not particularly

Table 3

Comparison of Rating Means by the Instructor/Judge and Rating Means by Debaters on Specific Criteria

For Affirmative Speeches

	Instructor/Judge Means	Debater Means	Significance
Analysis	3.32	3.54	NS
Reasoning	3.38	3.47	NS
Evidence	3.36	3.45	NS
Organization	3.80	3.59	NS
Refutation	3.35	3.49	NS
Delivery	3.53	3.54	NS

For Negative Speeches

	Instructor/Judge Means	Debater Means	Significance
Analysis	3.32	3.43	NS
Reasoning	3.41	3.48	NS
Evidence	2.57	3.08	$p < .001$
Organization	3.30	3.28	NS
Refutation	3.30	3.57	$p < .01$
Delivery	3.52	3.53	NS

surprising that in both cases the debaters rated their peers higher than the instructor/judge did. What is interesting is that only evidence and refutation in negative speeches show significant mean differences. Debaters and the instructor/judge apply different standards of evaluation for these two specific criteria in negative speeches. Differences in these standards could affect debaters' accuracy of judgment. An important observation is that although debaters tend to rate negative speeches higher on these criteria, nevertheless correlations on these same criteria for the negative speeches are higher than other correlations for specific criteria. This suggests that although their standards may differ from the instructor/judge's, debaters still show a moderate degree of accuracy judging relative quality on these criteria. These results challenge Hill's explanation that judgment correlations are higher in the classroom than in tournament debating because in the classroom the instructor's standards are more obvious. In the present sample the opposite appears to be true that for criteria for which no difference in ratings between the debaters and the instructor/judge can be demonstrated debaters' judgments are less accurate. Merely making the standards of the instructor more obvious to debaters will not improve their accuracy of judgement. What debaters need is more ability to conceptualize at higher levels of cognitive learning.

Summary and Conclusions

Understanding how debaters' judgments of debate speeches compare to a judge's evaluation has implications for debate pedagogy. A review of previous research on the issue led to several hypotheses not directly tested before. Judgments by classroom debaters were compared to the evaluations by an instructor/judge in 100 two person debates. Results show that less ego involved debaters in the classroom do not judge the outcome of a debate any more accurately than other debaters. Both overall ratings and ratings on specific criteria by less ego involved debaters in the classroom show significant moderate correlations with the judge's ratings. But these correlations are not high enough in comparison with other studies to conclude that debaters' judgments markedly improve when they are not ego involved in the outcome of the debate or are judging classroom rather than tournament debates. Apparently the side of the speech being evaluated affects the accuracy of debaters' judgements. Correlations of overall and specific criteria ratings are higher for negative debate speeches than for affirmative. Debaters have different standards for evidence and refutation in negative speeches than the instructor/judge has.

These results confirming those of earlier studies lead to the general conclusion that student debaters do not adequately recognize quality in debates. The results tend to bear out the concern that while present practice of debate pedagogy may train skillfull debaters, it apparently does not sufficiently teach higher level cognitive awareness to enable debaters to evaluate the quality of a debate.

More attention should be given in the debate classroom to listening, critical awareness of standards and conceptualization. Debaters need practice in evaluating debate and explaining the basis of their judgements in order to develop Bloom's higher level cognitive skills. One obvious learning exercise for the debate classroom is to debate evaluations of other debates. Debaters could debate among themselves or with the instructor/judge the ratings each assigned to a particular debate.

It would be useful also in debate pedagogy to separate evaluation of negative speeches from evaluation of affirmative speeches. Apparently the process of judgments of the two sides of a debate are different for student debaters. Further research is needed to make clear exactly why this is so, but until this research is completed, debate instructors should remain aware that the side of the debate judged affects a debater's judgement of that debate.

More attention should be given also in the debate classroom to clarifying the meaning of particular criteria of judgement. Until this is done, debaters in the classroom as well as in tournaments, when judging the debates of others as well as their own debates, will continue to judge debates less accurately and with different standards than judges.

References were retyped for readability

REFERENCES

[1]Joseph Baccua, "Debaters Judge Each Other," Quarterly Journal of Speech, XXIII (1937), 74-80.

[2]Benjamin S. Bloom, ed., Taxonomy of Education Objectives, Handbook 1: Cognitive Domain (New York: David McKay, 1956).

[3]Thomas King and Theodore Cleavenger, "A Comparison of Debate Results Obtained by Participant and Critic Judging," Southern Speech Journal, XXV (1960), 223-32.

[4]Larry L. Barker, "A Comparative Analysis of Debater-Judge Ratings," Journal of the American Forensic Association, II (January, 1965), 17-20.

[5]Sidney R. Hill, Jr., "A Study of Participant Evaluations in Debate," Journal of the American Forensic Association, IX (Winter, 1973), 371-77.

[6]Hill, p. 377.

[7]Barker, p. 20.

[8]Judee K. Burgoon, "Evaluation Criteria As Predictors of Debate Success," Journal of the American Forensic Association, XII (Summer 1975), 1-4.

[9]Dominic A. Infante, "The Percieved Educational Value of the AFA Form C and Flow Sheet Debate Ballots," Journal of the American Forensic Association, VII (Spring 1971), 308-9.